Computing @ Fermilab

Amitoj Singh

July 23th 2013

Fun Facts

1950: Computer fits in a single room.

1960: Computer is moveable across the room.

1970: Computer is moveable by just one person.

1980: Computer fits on my desk.

1990: Computer fits in my lap.

2000: Computer fits in my pocket.

2010: I’ve lost my computer.

Computing @ Fermilab

Core Computing Sector

Service Desk

Computer Security

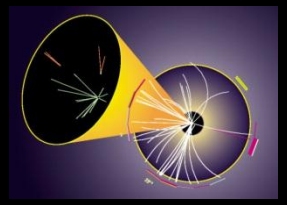
E - Communication

Science & Computing

Dzero

CMS

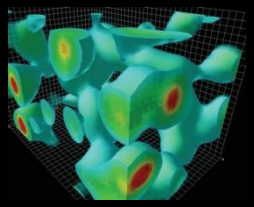
CDF



CDMS

MINOS

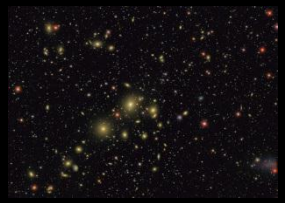
Lattice QCD



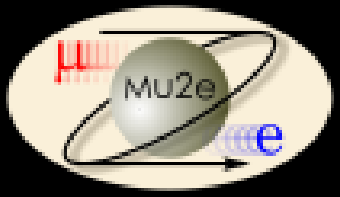
SDSS

NOvA

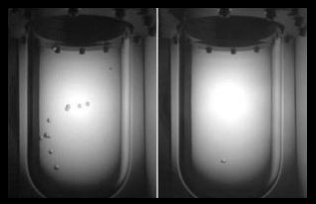
DES



Mu2e



COUPP



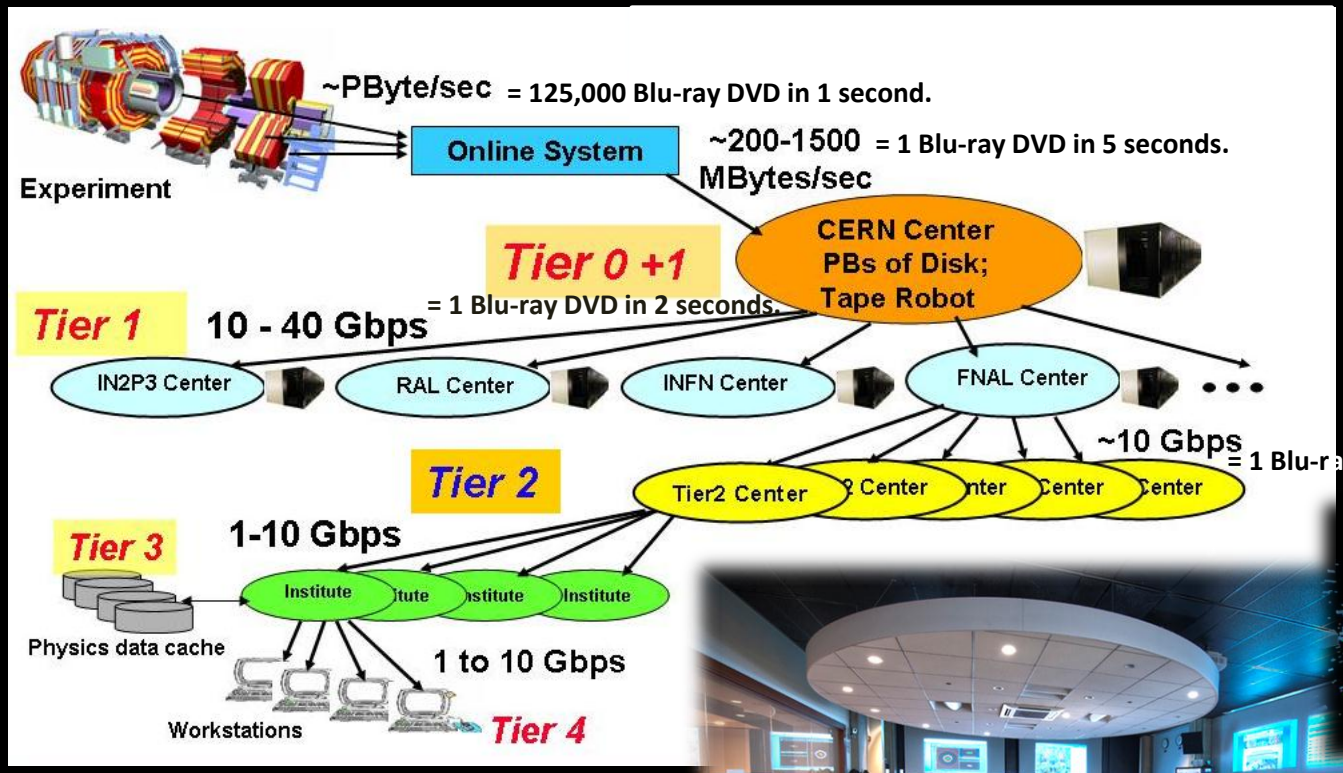
Pierre Auger

A major part of the Computing Sector's work is to support & improve the scientific programs at the lab. This includes computer support for experiment systems, design and implementation of the Data Acquisition and control systems, accelerator & detector simulations, research & development of the physics analysis software.

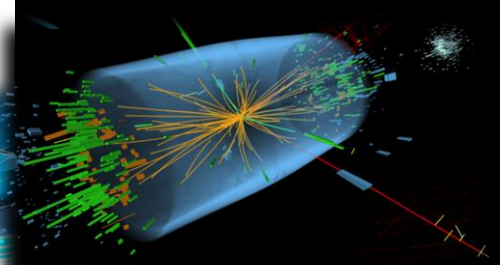
Computing | Data Handling & storage | Networking | Analysis Software

LHC CMS Experiment

one of many experiments supported by the Fermilab Computing Sector



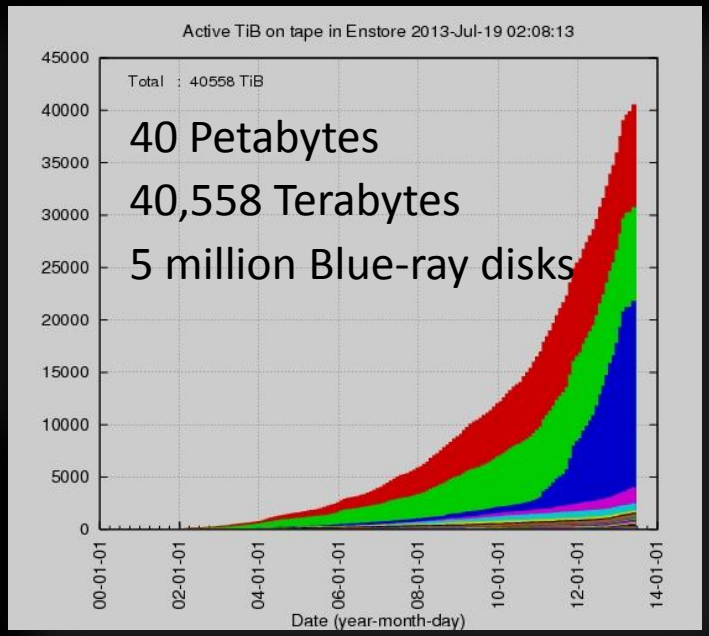
The distributed computing system for LHC distributes hundreds of Terabytes a day outward from CERN to the 11 distributed Tier-1 sites (one at Fermilab) & then 25 University Tier-2 and access to more than 100 smaller Tier-3 centers.



The CMS Remote Operations Center at Fermilab remotely supports the CMS experiment located 4,000 miles away in France. The ROC allows US physicists to help operate the CMS detector.

Data Handling & Storage

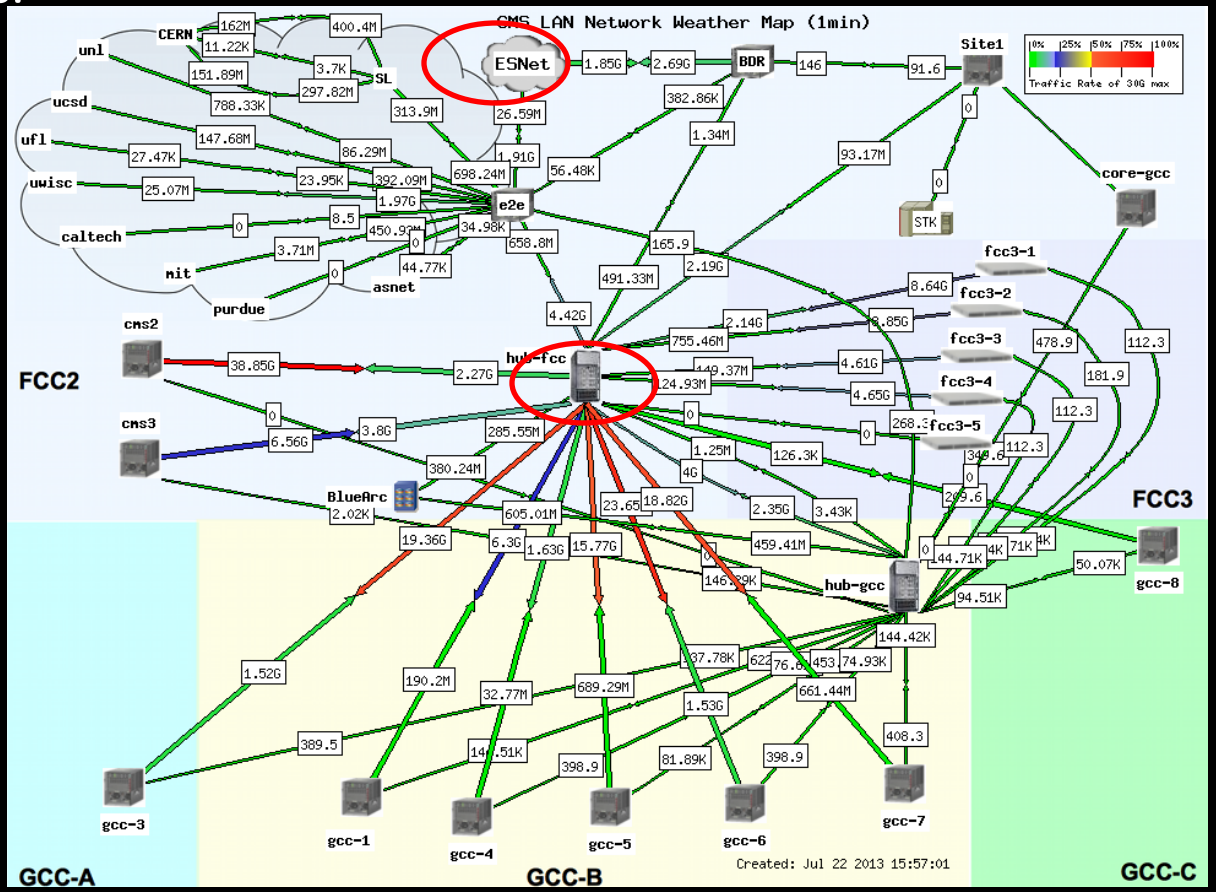
- **Enstore** (archival tape storage, 40 Petabytes stored, users transfer 3.2PB or 400,000 blue-ray DVD disks worth of data per day !!)
- **dCache** (100's of terabytes of disk front-end to Enstore for faster access)
- Databases (Payroll, Human Resources, Procurement, etc.)
- Disk and tape Backups for local and remote users.



T10KC Tape, storage capacity 5TB

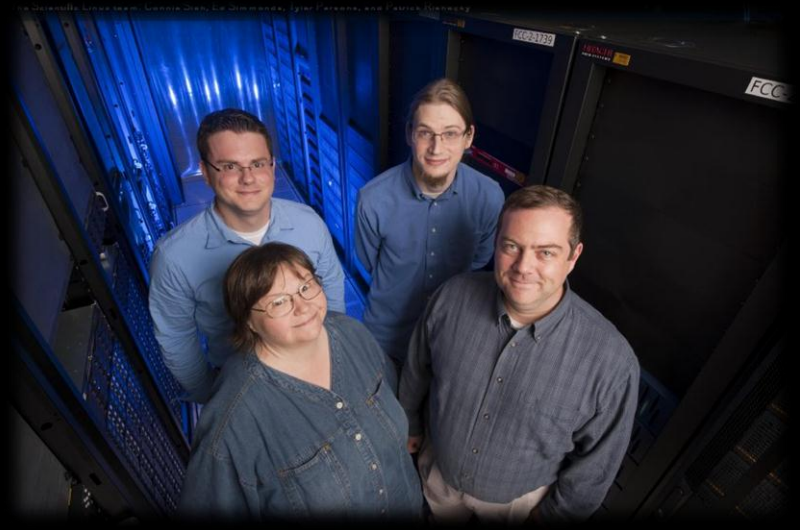
Networking

- Within Fermilab, users can transfer data at a rate of 8 gigabytes per second. That is equivalent to transferring an entire film on Blu-ray DVD every couple of seconds.
- By setting up dedicated pathways on which to transfer data to and from locations outside of the laboratory, Fermilab allows users to transfer data off-site at a rate of 10 gigabytes per second.



What is Scientific Linux?

- Created in 2004 at Fermilab, Scientific Linux is a Linux operating system distribution assembled by Fermilab and CERN in collaboration with other HEP institutions.
- 100% open source and free.
- Scientific Linux is used as the computing platform for major research projects all around the globe.
- Supported by an active user community.
- Packaged by a dedicated and professional team.



Switching Gears

Computing

High
Performance
Computing

Any questions thus far ?

What is HPC?

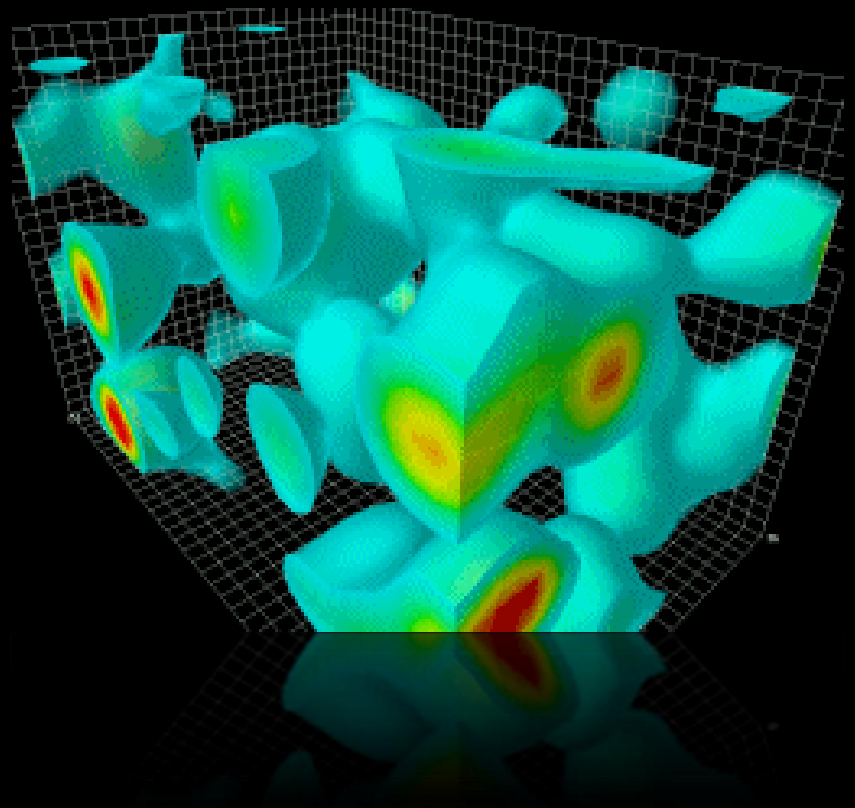
High Performance Computing (HPC) uses supercomputers and computer clusters to solve advanced computation problems.

Computer Cluster

IBM Blue Gene Supercomputer

Why we need HPC?

Discovered in the early 1970s, the theory of Quantum chromodynamics (QCD) consists of equations that describe the strong force that causes quarks to clump together to form protons and other constituents of matter. For a long time solving these equations was a struggle. But in the last decade using powerful supercomputers theorists are now able to finally solve the equations of QCD with high precision.



How do I measure the speed of a supercomputer?

FLOPS

FLoating point **O**perations **P**er **S**econd

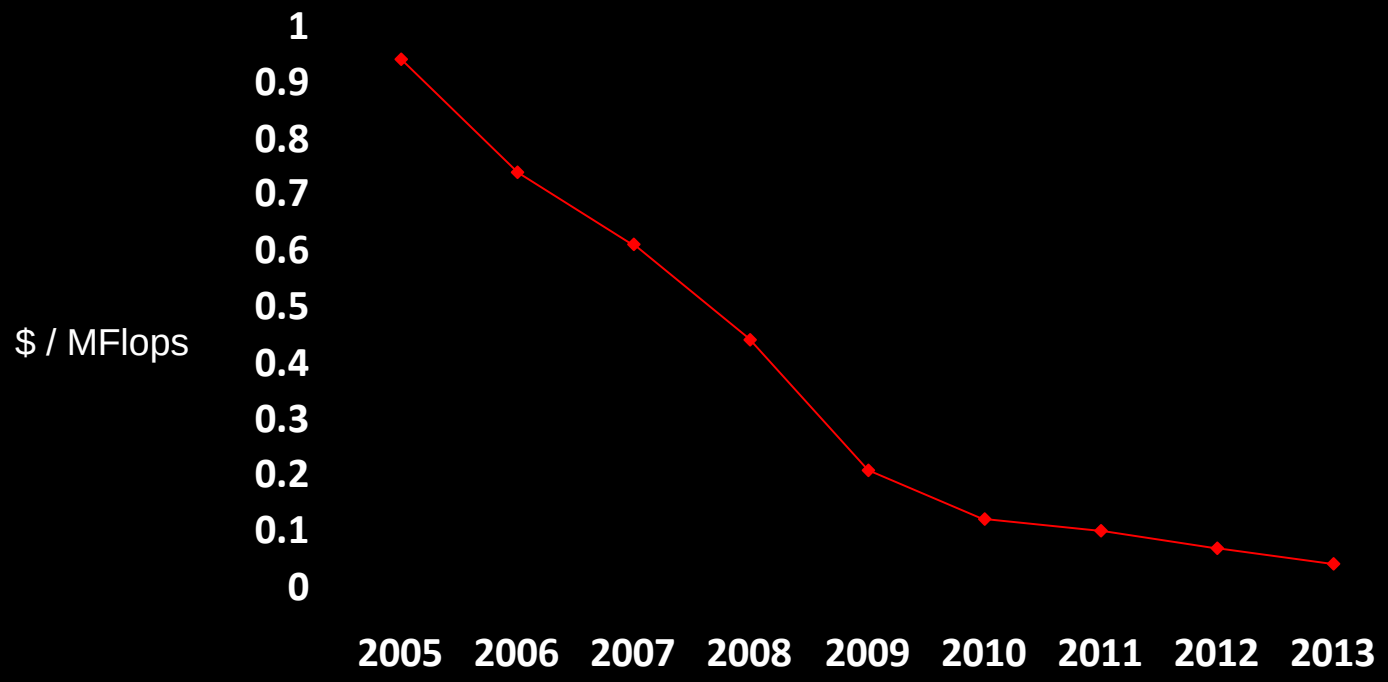
Examples of floating point numbers are

1.234567, 123456.7, 0.00001234567, 1234567000000000

LINPACK Benchmark

(http://www.top500.org/project/linpack/)

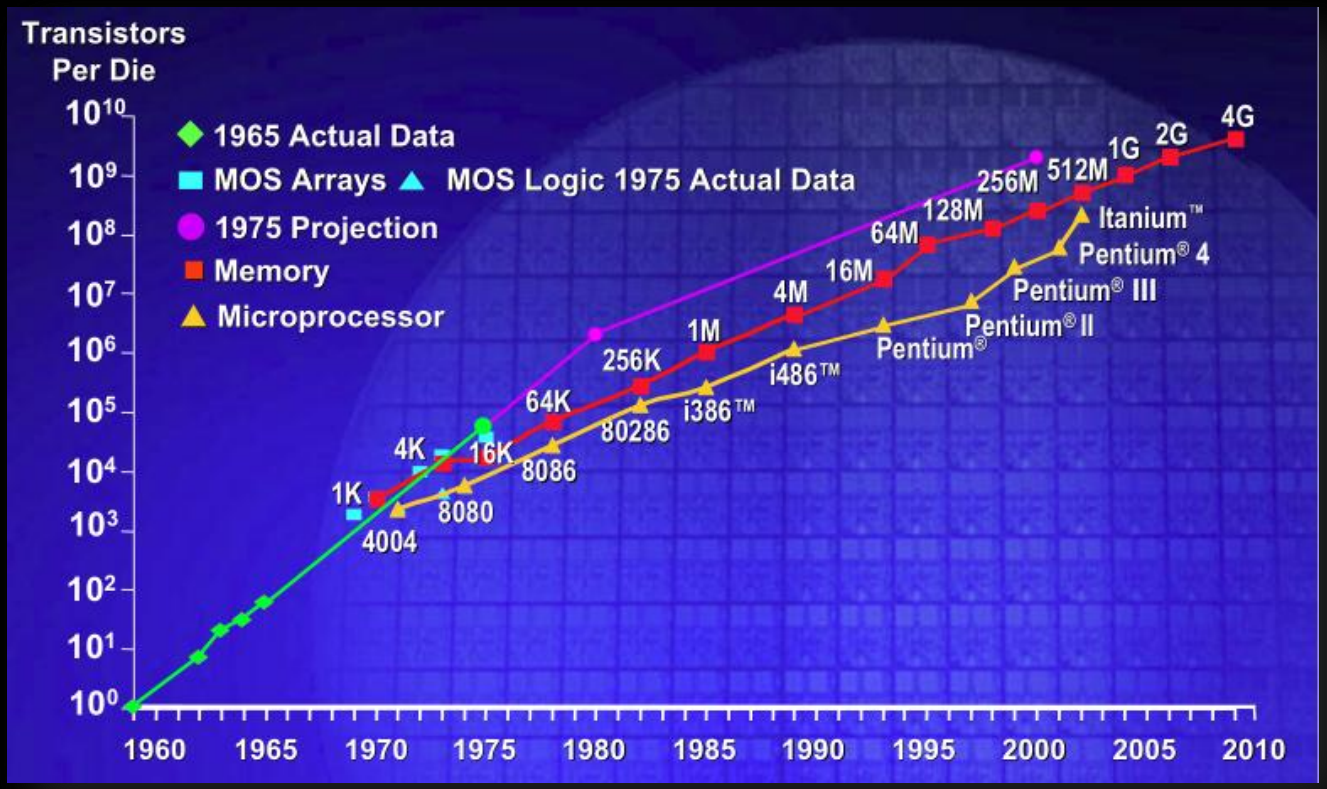
HPC cost trend



How much does 1 Million Flops cost?

HPC Industry Laws – Moore's Law

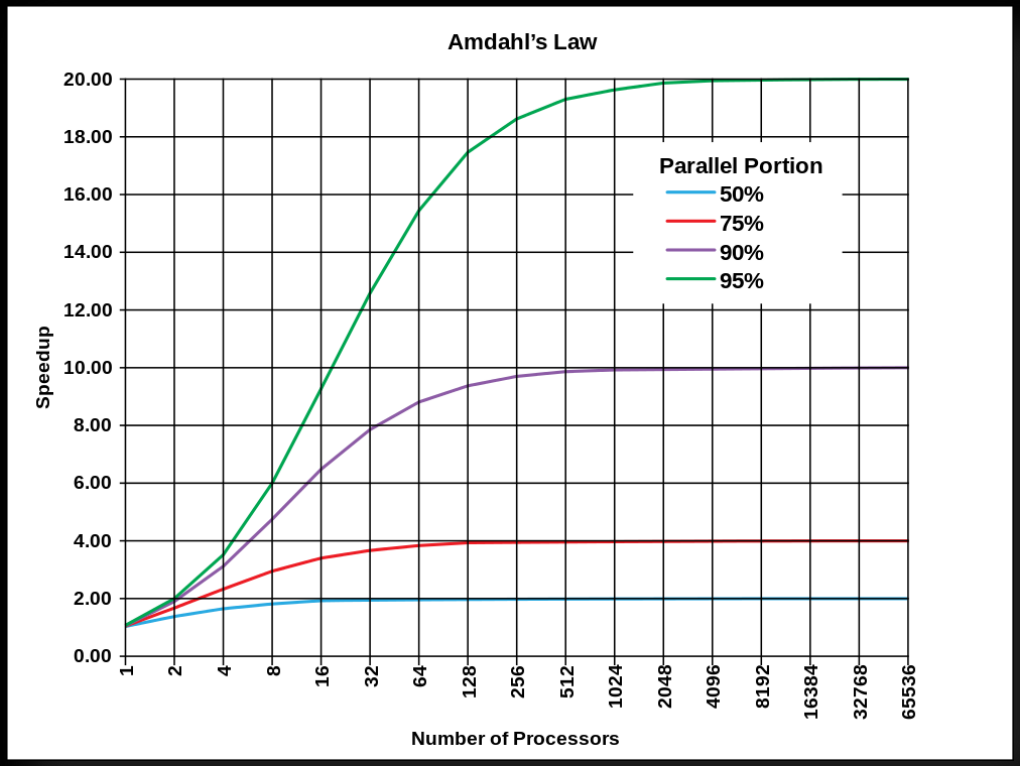
"The number of transistors that can be fabricated on a very large-scale integrated (VLSI) chip doubles every two years." - *Intel co-founder Gordon Moore* 1965



HPC Industry Laws – Amdahl's Law

"The speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program."

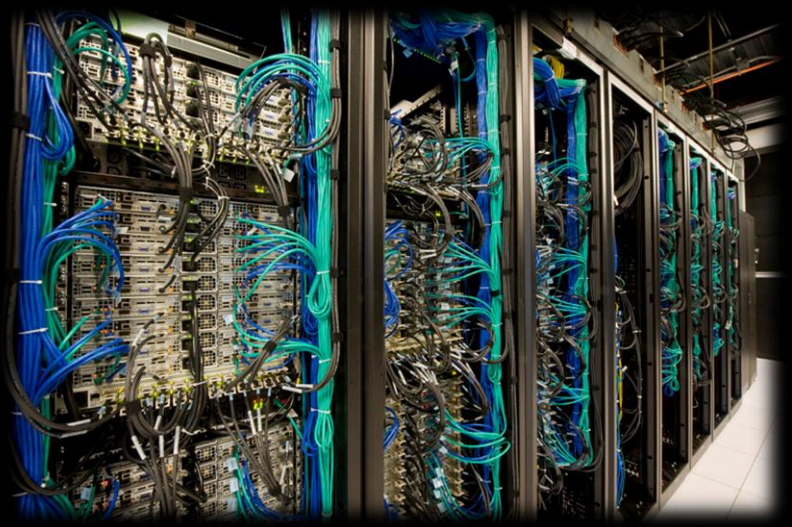
- *Computer architect Gene Amdahl, 1967.*



Computer Cluster Architecture

The building blocks of our computer clusters are:

- Compute nodes.
- Network switches.
- Lots of disk storage.



A typical compute node

Intel Xeon Phi Coprocessor

1,200,000 MFlops

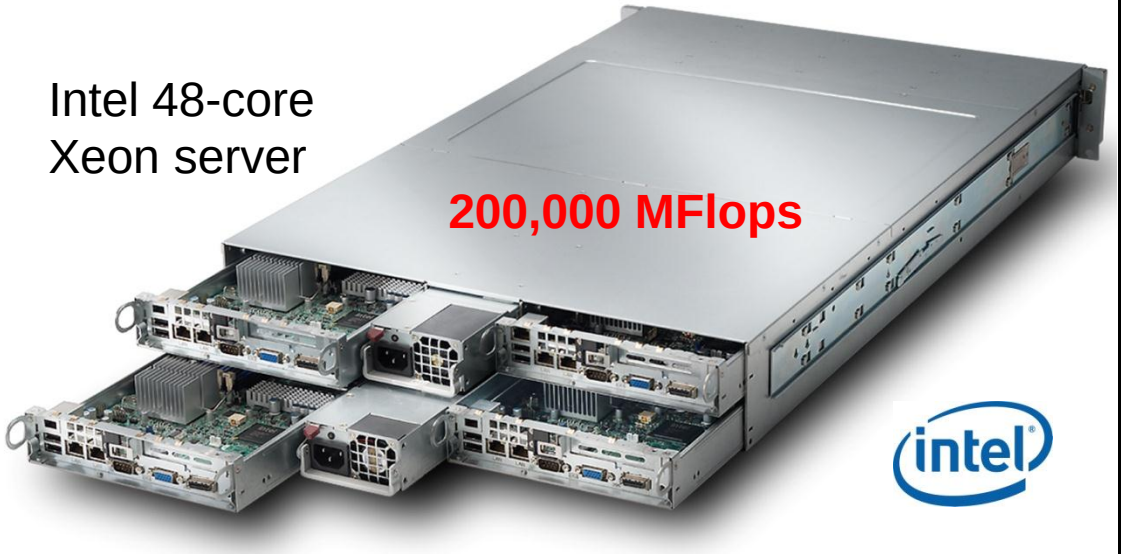
NVIDIA Tesla K20X GPU

1,310,000 MFlops

SONY

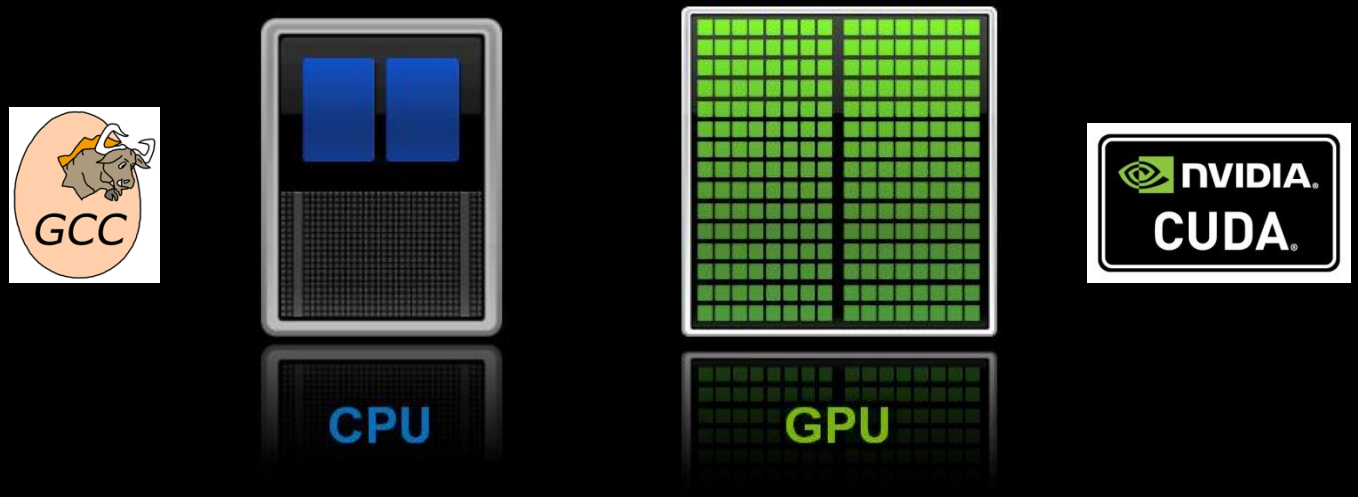
20,000 MFlops

Intel 48-core
Xeon server



200,000 MFlops

Difference between a CPU & GPU



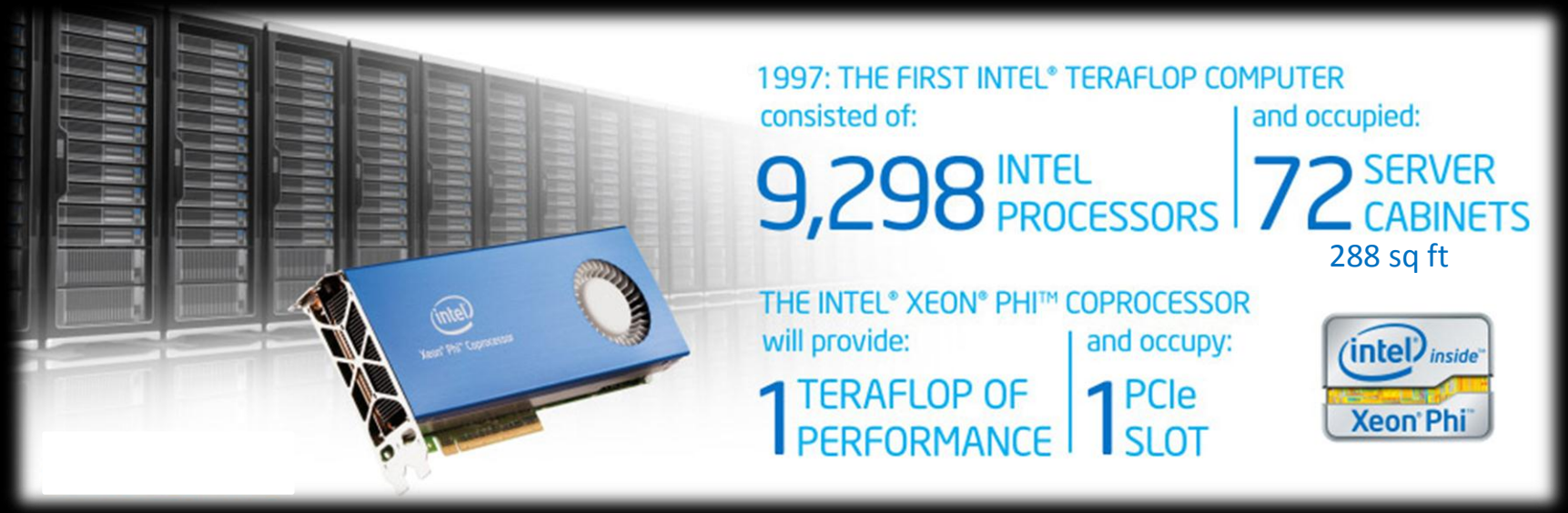
"GPUs are optimized for taking huge batches of data and performing the same operation over and over very quickly, unlike PC microprocessors, which tend to skip all over the place."

- Nathan Brookwood (Principal Analyst Insight64)

"The combination of a CPU with a GPU can deliver the best value of system performance, price, and power."

- Kevin Krewell (Senior editor Microprocessor Report)

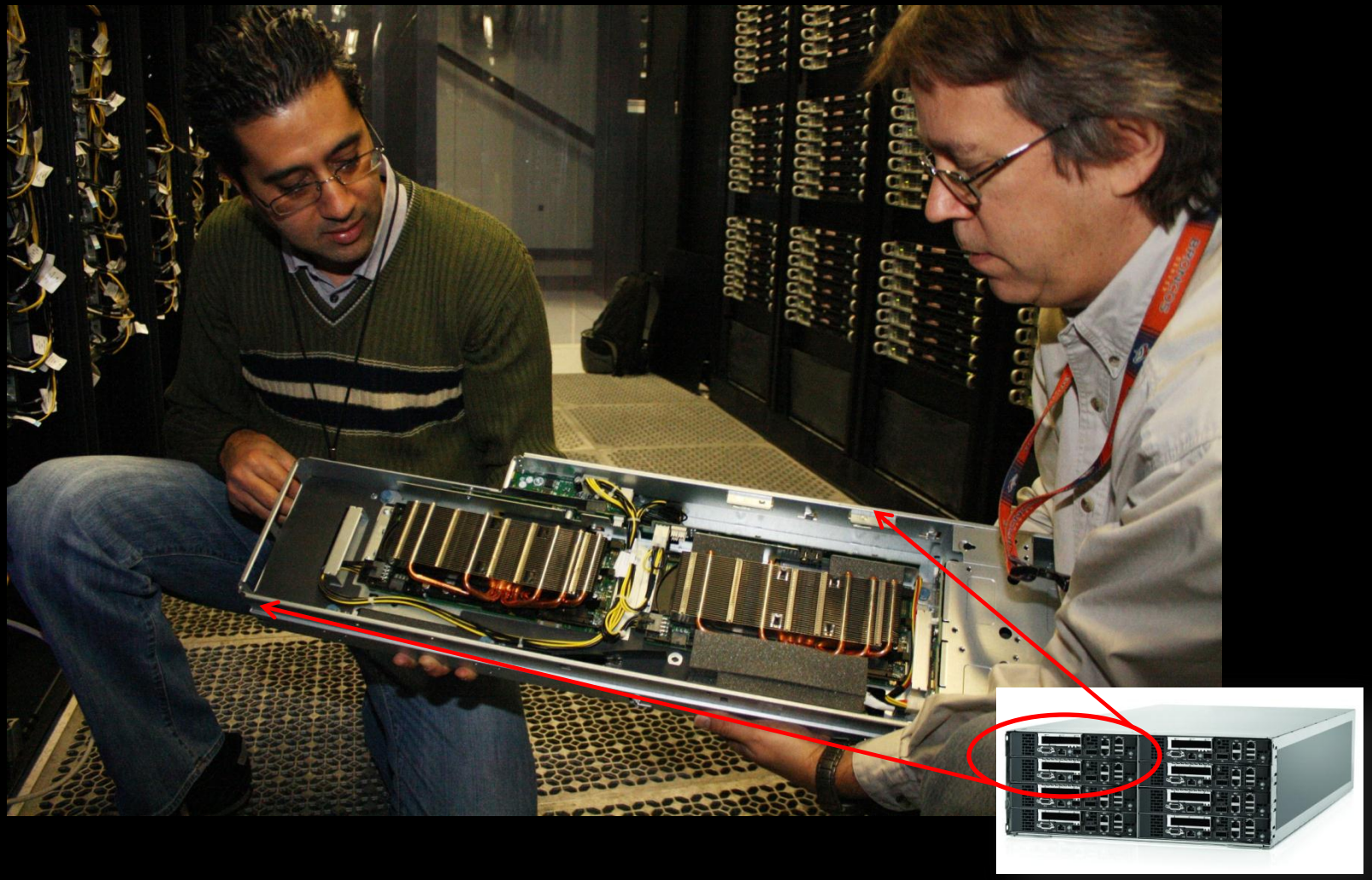
Intel Phi Co-processor



Graphic courtesy of Intel Corporation

- Intel's Phi co-processor is well-suited for workloads that are memory-bandwidth bound, such as Lattice QCD and memory-capacity bound, such as ray-tracing.
- Each coprocessor contains up to 61 cores, 244 threads and 16GB of GDDR5 memory (352 GB/s bandwidth).
- The coprocessor appears as an independent server and can run Scientific Linux while consuming as low as 225 Watts. A typical CPU-based server consumes about 600 Watts !!.

GPU: Graphics Processing Units



Networking: Bandwidth v/s Latency

When selecting network switches for supercomputers we have to consider two key factors: Bandwidth and Latency . . . . and price at times since some high speed switches can be prohibitively expensive.

How much can you carry?

How fast can you carry it?

Network Switches

This refrigerator size network switch built by Sun Microsystems consists of 3,456 ports and is capable of transferring 14 TBytes/second which is about 3000 DVDs worth of data in one second.

We use the smaller version of this switch on our Fermilab supercomputers.

User Interface to HPC

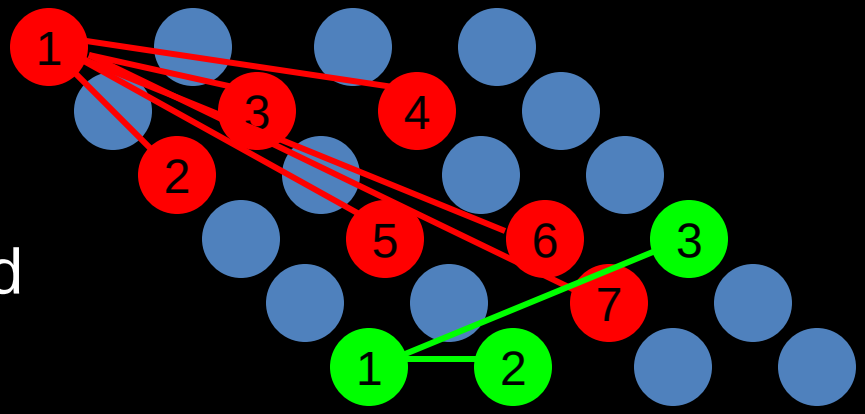
- The resource manager TORQUE maintains a *queue* of all such requests and assigns available and appropriate compute nodes to requests either FIFO (First In First Out) or depending on preset criteria.
- For example: TV screens that run commercials at gas stations or supermarket checkout lanes use the same concept. Frozen dinner entrée ads should only run after 6PM and cereal ads should run between 6 and 10AM. All other commercials are run FIFO (First In First Out).

Message Passing Interface

MPI is a language-independent communications protocol used to program parallel computers.

MPI's goals are high performance, scalability, and portability.

How do users figure out their resource (nodes, memory) requirements?



MPI

Argonne NATIONAL LABORATORY

Simple parallel code - an example

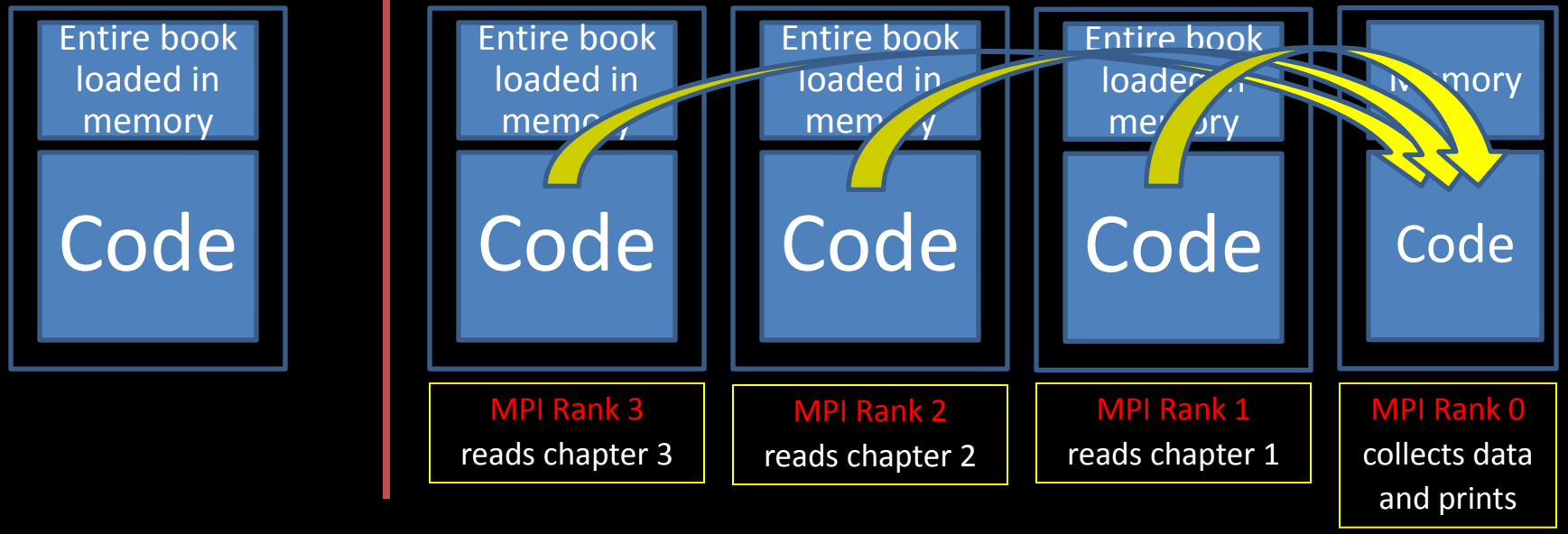
Serial Code

```
word_count = 0;

while (not end of book)
{
  function_readAword;
  word_count = word_count + 1;
}

print (word_count);
```

Serial to Parallel

Parallel Code run on four computers

```
word_count = 0;
myChapter = myMPIrank;
goto myChapter;

while (not end of chapter)
{
   function_readAword;
   word_count = word_count + 1
}
sendToMaster (word_count)
```



Managing Supercomputers

- <u>Biggest challenge:</u> A job on the supercomputer will run at the speed of the slowest component.

• 2.4GHz Intel Core 2 Duo processor with 3MB on-chip shared L2 cache	2,000 MFlops
• 32GB of 1066MHz DDR3 SDRAM	2,000 MTransfers/s
•1TB 5400-rpm Serial ATA hard disk drive	50 MBytes/s

The disk which is 40 times slower is the slowest component!!

Computing Facilities

Our esteemed users

Conclusion

It is an exciting time to be in the field of computing which is at it’s peak in terms of potential, available hardware and software options and the variety of research that can be conducted using the computing power provided by the world’s fastest custom or purpose built supercomputers.